Trabajo Estructuras Discretas II

Instrucciones:

- **El presente trabajo corresponde a la nota correspondiente al Permanente II**
- **Deberá hacer uso de lo aprendido en clases y de lo que ya conoce de Programación I: arrays, lectura de archivos, bucles, selección, llamada a funciones y demás**
- **El programa deberá ser remitido mediante el Moodle con fecha lunes 03 de diciembre del 2007 a horas 23:00**
- **Seguir las instrucciones sobre la realización del programa dado al final de ésta hoja.**
- **Se calificará la correctitud y simplicidad del código, así como los comentarios adecuados que se coloquen al interior del programa**
- **No existirá otra fecha de presentación de este programa más que la del día lunes 03 de diciembre, prever esto a fin de evitar inconvenientes**

Problema a ser resuelto:

La aplicación que desarrollará será una correspondiente a Bioinformática donde se requieren que grandes cantidades de DNA sean procesadas. Las cadenas de DNA pueden ser representadas como del tipo string y que contienen las siguientes 4 letras (en biología denominadas nucleótidos): A,C,T y G. Por ejemplo:

- ACTTGGGAATTTAACCCCCCCAAAACCCCCC
- GGGGGGGAATCCCCACCCCTCCCCACCGGATTT

Algunas veces un biólogo obtendrá la secuencia de un segmento de DNA de algún organismo y luego querrá saber si algo similar ha sido encontrado antes. Una búsqueda será realizada en la base de datos para algo similar.

¿Qué significa similar? Se necesita el definir una función de distancia entre dos cadenas, escoger un umbral, y luego ver si las dos cadenas están por debajo de éste umbral diríamos que son similares. Muchas funciones de distancia han sido propuestas, nosotros usaremos una denominada d^2.

El d^2 trabaja en frecuencias de palabras. Para cada cadena construimos una tabla que nos diga para cada palabra en la cadena cuantas veces ésta aparece (en este trabajo se considerarán palabras de longitud 3). Imaginemos que la palabra w aparece $c_x(w)$ veces en las cadena x y $c_y(w)$ veces en la cadena y. Existen 64 palabras distintas de longitud 3 (del alfabeto A,C,T,G). Denotemos a este conjunto W. Luego el d^2 calcula el puntaje entre x e y definido como sigue:

$$d^2(x,y) \overset{def}{=} \sum_{w \in W} (d_x(w) - d_y(w))^2$$

Ejemplo:

x=ACTTGTTG
y=ACTTACT

Las palabras que aparecen en x son: ACT, CTT, TTG, TGT, GTT. Las palabras que aparecen en y son: ACT, CTT, TTA, TAC, y la tabla de frecuencias de palabras es:

Palabra	x	y	Dif	Cuad. Dif
ACT	1	2	-1	1
CTT	1	1	0	0
GTT	1	0	1	1
TGT	1	0	1	1
TAC	0	1	-1	1
TTA	0	1	-1	1
TTG	2	0	2	4

Por tanto, $d^2(x,y)=9$

Hint: Sería conveniente tener una matriz que represente las tablas de frecuencias, con una entrada para cada palabra. Podría ser un array indexado por un entero positivo. Existen 64 (4^3) palabras distintas de longitud 3, así de que una solución sería el representar cada palabra como un único número en el rango de 0…63. Esto puede ser hecho al asignar a las letras valores enteros: A=0; C=1; G=2; T=3. Entonces una palabra como $s_0s_1s_2$ puede ser representada por el número $\delta(s_0)+4\delta(s_1)+16\delta(s_2)$

Programa
Su tarea consistirá en escribir un programa en Java que lea dos cadenas desde un archivo y que calcule el puntaje d^2 entre ambas. Puede asumir que los datos de entrada no contienen errores.
También se pide que el programa calcule el complemento inverso de ambas cadenas, esto consiste en sustituir la A por la T y la C por la G. Este complemento deberá ser hallado al comparar x con el complemento inverso de y.

Instrucciones de Entrega:

- El programa deberá ser llamado D2.java
- El formato de la salida será el siguiente:
  Primera secuencia en una línea
  Segunda secuencia en una línea
  Complemento inverso de la segunda secuencia
  Puntaje d2 entre la primera y segunda secuencia en una línea
  Puntaje d2 entre la primera secuencia y el CI de la segunda

Trabajo de Investigación:
Instrucciones:

- **Fecha de entrega igual que el trabajo de programación**
- **Hacer uso de la plantilla para Word de ACM**
- **Se revisarán calidad del artículo así como el eficiente uso de referencias cruzadas**
- **El artículo deberá de constar de 6 a 9 páginas como máximo**

Tema: Estado del arte del algoritmo del Stable Marriage